同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

交通数据分析

第四讲 广义线性回归模型

沈煜 博士 副教授
嘉定校区交通运输工程学院311室
yshen@tongji.edu.cn
2022年03月18日

计划进度

周	日期	主讲	内容	模块
1	2022.02.25	沈煜	概述	
2	2022.03.04	沈煜	在线数据采集方法	爬虫
3	2022.03.11	沈煜	线性回归模型	回归分析
5	2022.03.18	沈煜	广义线性回归	
4	2022.03.25	沈煜	广义线性回归（作业1）	
6	2022.04.01	沈煜	空间数据描述性分析	
7	2022.04.08	沈煜	空间自回归方法（作业2）	
8	2022.04.15	沈煜	关联：Apriori	机器学习
9	2022.04.22	沈煜	决策树、支持向量机（作业3）	
10	2022.04.29	沈煜	浅层神经网络	
11	2022.05.06	沈煜	卷积神经网络（期末大作业）	
12	2022.05.13	沈煜	经典网络结构	
13	2022.05.20	沈煜	聚类：K-Means、DBSCAN	
14	2022.05.27	沈煜	贝叶斯方法、卡尔曼滤波	
15	2022.06.03	-	端午节放假	
16	2022.06.10	沈煜	期末汇报（1）	
17	2022.06.17	沈煜	期末汇报（2）	

主要内容

- 线性回归：统计检验
- 广义模型基础
- 逻辑回归

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

统计检验

小结

- **回归与回归方程**
  - 线性回归模型形式
  - 模型理论假设
- **参数估计**
  - 最小二乘法
  - 极大似然估计
- **残差分析**
  - 线性、正态、方差齐性
- **统计检验**
  - 变量的显著性检验
  - 拟合优度检验

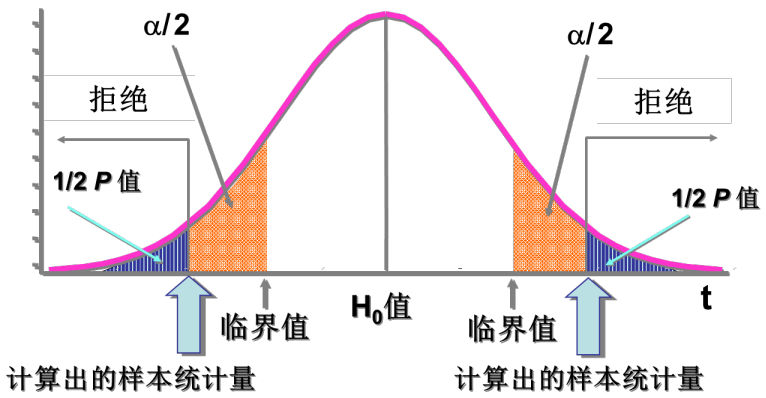
变量的显著性检验

- 回归分析是要判断解释变量x是否是被解释变量y的一个显著性的影响因素,即需要进行变量的显著性检验
- 变量的显著性检验所应用的方法是数理统计学中的假设检验
- 计量经计学中，主要是针对变量的参数真值是否为零来进行显著性检验的

显著性检验

（1）对总体参数提出假设

H_0：$\beta_1=0$，　H_1：$\beta_1 \neq 0$



（2）以原假设H0构造t统计量

$$t=\frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

（3）给定显著性水平α，查t分布表，得临界值$t_{\alpha/2}(n-2)$

(4) 比较，判断

若　$|t|> t_{\alpha/2}(n-2)$，则拒绝H_0，接受H_1；

若　$|t|\leq t_{\alpha/2}(n-2)$，则拒绝$H_1$，接受$H_0$；

拟合优度检验

拟合优度检验：对样本回归直线与样本观测值之间拟合程度的检验

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2$$

TSS = ESS + RSS

TSS--总离差平方和
RSS--回归平方和
ESS—残差平方和

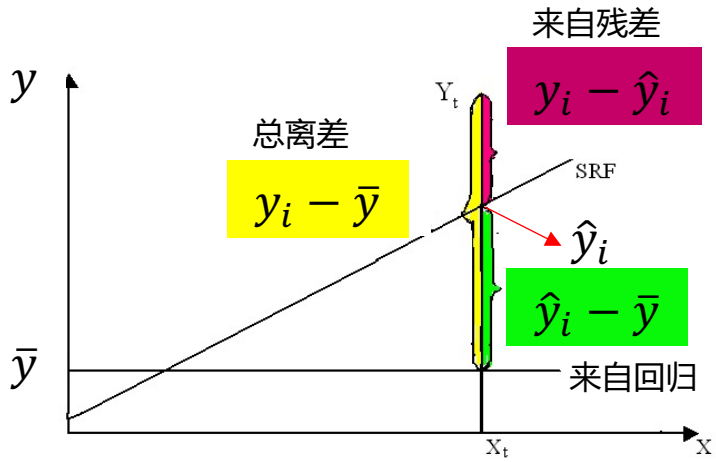
可决系数（判定系数）R^2 为:

$$R^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}=1-\frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}$$

范围[0,1]
越靠近1，模型对数据的拟合程度越好

拟合优度检验

➢$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = e_i + (\hat{y}_i - \bar{y})$



如果$y_i = \hat{y}_i$即实际观测值落在样本回归“线”上，则拟合最好。可认为，“离差”全部来自回归线，而与“残差”无关。

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

广义线性模型基础

线性回归

- $y = h(x) = \beta^T x$
  - 通过观察x，以一个简单的线性函数$h(x)$来预测y
- 因变量/响应变量y：预测的目标
  - 分布：在建模时，我们实际上关注的是当给定数据x和参数β时，$y|x,\beta$服从的分布。线性回归时，是正态分布。
  - 观察：观察的是采样，即真正观察到的结果，是一个值
  - 预测：预测的是期望。我们用$y = E[y|x] = h(x)$表示模型的预测。虽然y实际上服从一个分布，但是预测的结果是整个分布的均值μ，它只是一个值。
- 自变量x：特征，predictor
- 假设：$h(x) = \beta^T x$
  - 系数β与特征x的内积(inner product)，表示的是线性特征：广义线性模型基于此的推广
  - 特征x_j通过系数β_j线性加和，不同系数β_j反映不同x_j对y的贡献程度

线性模型的不足

- 假设因变量为连续的：实数域$(-\infty, +\infty)$
- 交通中，很多因变量均为非负
  - 比如事故数、排队长度
- 正态假设在实践中也许并不满足
- 同方差的假设，且方差不是其期望值的函数
  - 泊松分布：$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$
  - 期望值和方差都是μ
  - 二项分布：$f(y|p) = p^y(1-p)^{(1-y)}$
  - 期望值np和方差$np(1-p)$
- 解释变量只能通过加法对因变量产生影响

广义线性模型

- Generalized Linear Model
- Wiki定义：
  - In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.
- 三个部分：
  - 随机分布（Random component）
  - 系统模型（Systematic component）
  - 连接函数（Link function）

随机分布

- 响应函数y的分布必须服从某一指数家族的分布
- $y|x,\beta \sim Exponential\ Family(\eta)$
  - η是指数家族分布的自然参数
- 指数分布家族：
  - 线性回归服从高斯（正态）分布
  - 逻辑回归logistic regression服从伯努利分布
  - 还有多项式分布、拉普拉斯分布、泊松分布等等
- 也被称为误差结构Error structure
  - 线性回归的残差$\epsilon = y - h(x)$，服从高斯分布$N(0,\sigma)$
  - 如果没有直接的误差项（如逻辑回归），可构造其它的残差（如服从二项式分布binomial）

系统模型

- 广义线性模型本质上还是线性模型
- 推广的是响应变量y的分布
- 模型最终还是去学习/拟合$\beta^T x$中的系数β
- GLM里，假设的是$\eta = \beta^T x$
- y相关的指数家族分布里的自然参数η等于线性回归的值

连接函数

- 连接函数：连接了响应变量y和线性回归值$\eta = \beta^T x$
- 随机分布部分和系统模型部分把y和$\beta^T x$统一到指数家族分布中
- 接下来就是通过连接函数建立两者的联系
- 对任意指数家族分布，都有连接函数$g(\mu) = \eta$
  - μ是分布的均值
  - η是指数分布的自然参数
- 比如：
  - 线性回归（高斯分布）的连接函数就是自身
  - 伯努利的连接函数是logit函数：$g(\mu) = \ln \frac{\mu}{1-\mu} = \eta$

连接函数

- 连接函数建立了响应变量y的分布均值（就是回归出的目标）和线性回归的值之间的关系
- 连接函数的反函数是响应函数response function
- $g^{-1}(\eta) = \mu$
- 响应函数把线性回归的值直接映射到了预测目标y
- 比较常见的响应函数比如logistics函数就是logit的反函数

线性回归与广义线性模型

	线性回归	广义线性模型
线性回归值	$\eta = \beta^T x$	$\eta = \beta^T x$
响应变量分布	$y \sim N(\eta, \sigma^2)$	y~指数家族
连接函数	$\eta = g(\mu) = \mu$ Identity函数	$g(\mu)$ logit，伯努利等
预测值	$h(x) = E[y\|x, \beta]$ $= \mu = g^{-1}(\eta)$ $= \mu$	$h(x) = E[y\|x, \beta]$ $= \mu = g^{-1}(\eta)$ logistics

常用的广义线性模型

- 由先验信息选择分布类型
  - 常数方差→正态分布
  - 方差等于均值→泊松分布
  - 方差等于均值的平方→伽马分布
  - 方差等于均值的三次方→逆高斯分布

- 常用GLM
  - 逻辑回归
    - 0-1二元因变量
    - 编码为0和1完全是随意的，一般而言，关注的事件编码为1
  - 泊松回归/负二项回归
    - 因变量是计数数据，给定的一段时间内，一个特定时间发生的次数

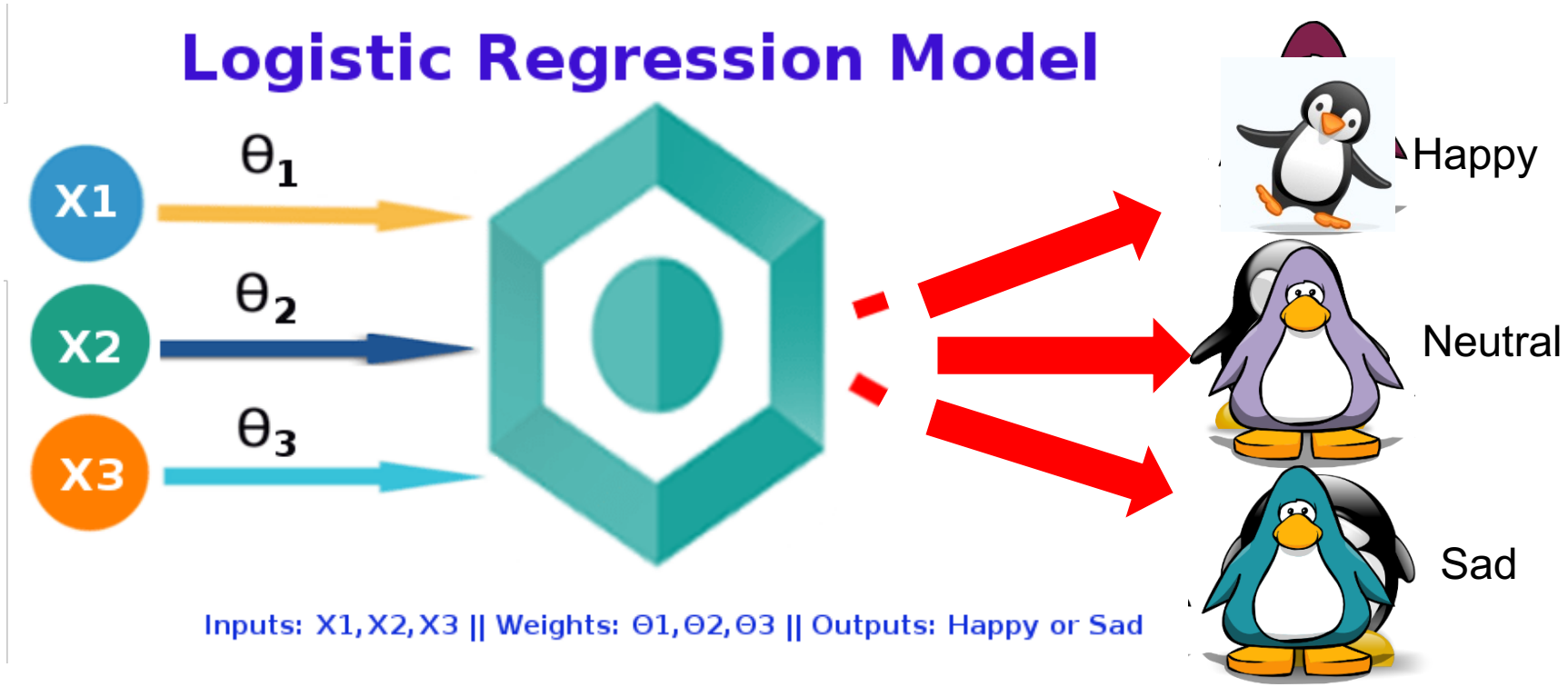
同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

逻辑回归

Logistic Regression

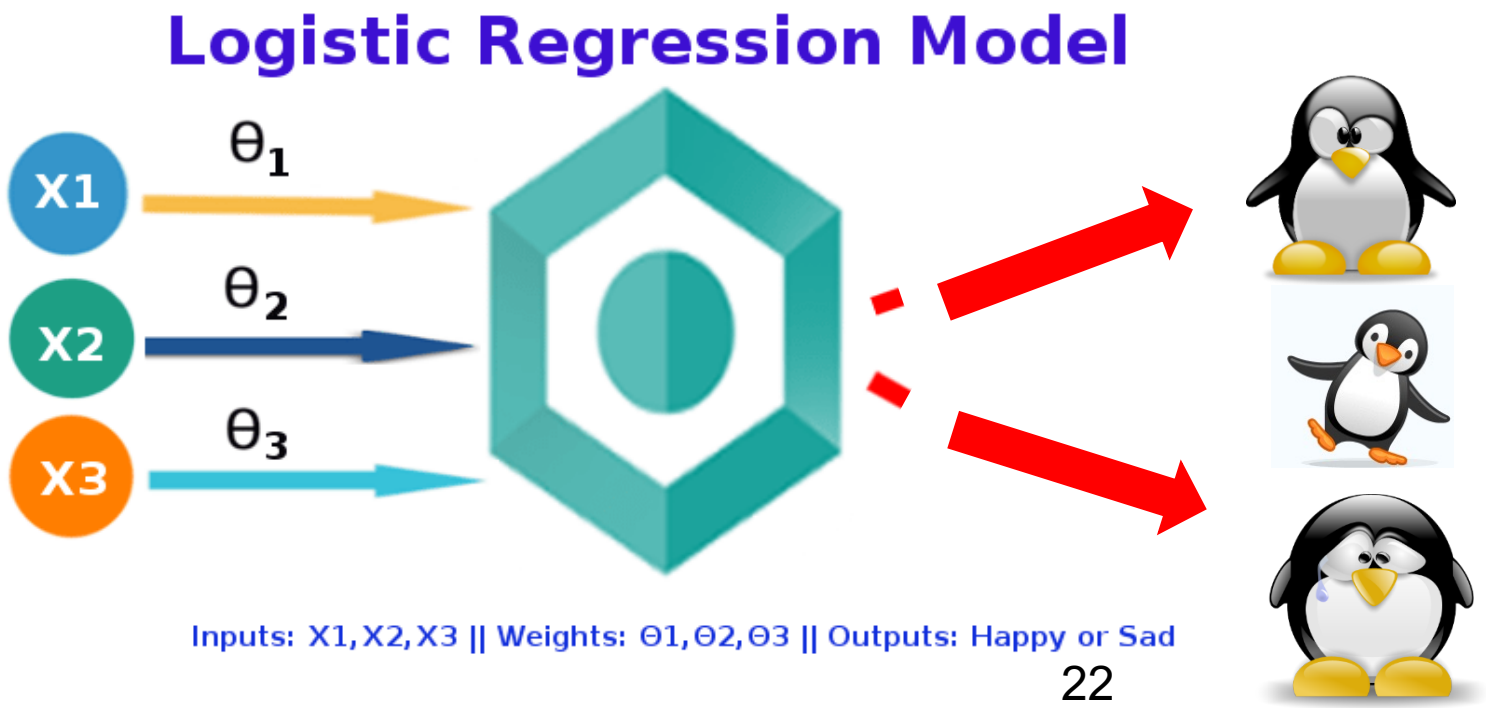
逻辑回归模型分类

- 按因变量分类
  - 二分类逻辑回归模型（ binary logistic regression model ）
  - 多分类无序逻辑回归模型 (Nominal logistic regression model)



逻辑回归模型分类

- 按研究设计分类
  - 非配对设计：非条件逻辑回归模型 (case-control logistic regression model)
  - 配对设计：条件逻辑回归模型 （ matched case-control logistic regression model ）



逻辑回归

- GLM中最常用的是逻辑回归，即y服从伯努利分布
- 随机分布：指数家族$y_i \sim Bern(p_i)$
- 线性回归值：$\eta_i = \sum_j^J \beta_j x_{ij}$，此部分GLM都有且一致
- 连接函数：$\eta = g(\mu) = \ln \frac{\mu}{1-\mu}$
  - logit函数，也称log-odds函数
- 响应函数：$\mu = g^{-1}(\eta) = \frac{1}{1+\exp(-\eta)}$
  - logistic函数，也称sigmoid函数
- 预测值：$h(x_i) = E[y_i|\beta, x_i] = logistic(\eta_i)$
- 损失函数：$E = -y \ln h(x) - (1-y) \ln\big(1 - h(x)\big)$

伯努利分布

- 分布律：$p(x|\mu) = Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$
- 写成指数分布家族的形式：
- $p(x|\mu) = \exp(\ln(\mu^x(1-\mu)^{1-x}))$
- $= \exp(x\ln\mu + (1-x)\ln(1-\mu))$
- $= \exp(x\ln\mu - x\ln(1-\mu) + \ln(1-\mu))$
- $= (1-\mu)\exp\left(x\ln\frac{\mu}{1-\mu}\right)$
- 得到的自然参数是连接函数
- logit函数：$\eta = \ln\frac{\mu}{1-\mu}$

逻辑回归模型

- 主要应用在研究某些现象发生的概率p，比如事故是否会发生，拥堵是否会发生，以及讨论概率p与哪些因素有关
- 作为概率值，一定有$0 \le p \le 1$（=0或者=1，是我们要研究的吗？）

- 不直接研究p，而是研究p的一个单调函数
- $\mathrm{Logit(p)} = \ln\frac{p}{1-p}$
- 当p从0→1，logit(p)的范围？

$$\theta = \ln\frac{p}{1-p} = \log it(p) \qquad \theta = \beta_0 + \beta_1 x + \varepsilon$$

- $\mathrm{Logit(p)} = \ln\frac{p}{1-p} = \beta_0 + \beta X + \varepsilon$
- $\mathrm{p} = \frac{\exp(\beta_0+\beta X+\varepsilon)}{1+\exp(\beta_0+\beta X+\varepsilon)}$

逻辑回归模型定义

$$\theta = \ln\frac{p}{1-p} = \log it(p)$$

Odds，比值/优势

$$OR = \frac{Odds_1}{Odds_2}$$

Odds Ratio，比值比/优势比

- $\text{Logit(p)} = \ln\frac{p}{1-p} = \beta_0 + \beta X + \varepsilon$
  - 当x=b时，logit(p)=$\beta_0 + \beta b$
  - 当x=a时，logit(p)=$\beta_0 + \beta a$
  - logit(p) at $_{x=a}$减去logit(p) at $_{x=b}$=（a-b）β
  - x=a相对于x=b的优势比是exp((a-b)β)

逻辑回归模型定义

- 当x是数值型变量时
  - Log(odds)增加β，当x增加1时
  - x=k+1相对于x=k，odds是exp（β）倍
- 当x是分类变量时
  - β=log(odds)|分类变量值为1-log(odds)|分类变量值为0
  - x=1相对于x=0，odds是exp（β）倍

逻辑回归模型定义

	发生事故	不发生事故
路面潮湿	15	50
路面干燥	20	200

在路面潮湿的情况下，事故发生几率p=15/65，事故发生优势p/(1-p)= 0.3
在路面干燥的情况下，事故发生几率p=20/220，事故发生优势p/(1-p)= 0.1

路面潮湿情况下与路面干燥情况下的优势比OR= 3

$$p = \frac{\exp(\beta_0 + \beta X)}{1 + \exp(\beta_0 + \beta X)}$$

Odds=$\frac{p}{1-p}$= $\exp(\beta_0 + \beta X)$

Odds Ratio=$\frac{Odds1}{Odds2}$= $\exp(\beta(X_1 - X_2))$

β = 1.0986

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

第四讲 结束